35.14356 PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	)	
	:	Examiner: NYA
NORIKI OTANI ET AL.	)	
	:	Group Art Unit: 2776
Application No.: 09/533,255	)	
	:	
Filed: March 23, 2000	)	
	:	
For: APPARATUS AND METHOD FOR DIVIDING DOCUMENT INCLUDING TABLE	)	July 27, 2000

Assistant Commissioner for Patents
Washington, D.C. 20231

CLAIM TO PRIORITY

Sir:

Applicants hereby claims priority under the International Convention and all rights to which they are entitled under 35 U.S.C. § 119 based upon the following Japanese Priority Applications:

11-077583 Filed March 23, 1999
2000-081870 filed March 23, 2000

Certified copies of the priority documents are enclosed.

Applicants' undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our address given below.

Respectfully submitted,

Jack M. Arnold

Attorney for Applicants

Registration No. 25,823

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 99071 v 1

CTO 1755645

09/530,755

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年　３月２３日

出 願 番 号
Application Number: 平成１１年特許願第０７７５８３号

出 願 人
Applicant (s): キヤノン株式会社

２０００年　４月１４日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆 彦

出証番号　出証特２０００－３０２７０５８

【書類名】　特許願

【整理番号】　3931016

【提出日】　平成11年 3月23日

【あて先】　特許庁長官　伊佐山　建志　殿

【国際特許分類】　G06F 17/30

【発明の名称】　文書分割装置及び方法、及びそのプログラムを記憶した記憶媒体

【請求項の数】　15

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　大谷　紀子

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　藤井　憲一

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　伊藤　史朗

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　上田　隆也

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【氏名】　池田　裕治

【特許出願人】

【識別番号】　000001007

【住所又は居所】　東京都大田区下丸子３丁目３０番２号

【氏名又は名称】　キヤノン株式会社

【代表者】　御手洗　冨士夫

【電話番号】　03-3758-2111

【代理人】

【識別番号】　100069877

【住所又は居所】　東京都大田区下丸子３丁目３０番２号キヤノン株式会社内

【弁理士】

【氏名又は名称】　丸島　儀一

【電話番号】　03-3758-2111

【手数料の表示】

【予納台帳番号】　011224

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9703271

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書分割装置及び方法、及びそのプログラムを記憶した記憶媒体

【特許請求の範囲】

【請求項１】　処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析手段と、

該テーブル解析手段により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定手段と、

前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成手段と、

前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成手段とを備えたことを特徴とする文書分割装置。

【請求項２】　前記第１のセグメント生成手段が、

前記セル位置データおよび前記セルベクトルを参照して、前記テーブルにおいて各データが行または列のどちらで表現されているかを判別し、当該テーブルの分割方向を決める分割方向決定手段と、

前記テーブルタイプおよび前記分割方向を参照して、前記テーブルを分割してセグメントを生成する表セグメント生成手段とを備えたことを特徴とする請求項１に記載の文書分割装置。

【請求項３】　前記第２のセグメント生成手段が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項２に記載の文書分割装置。

【請求項４】　前記第２のセグメント生成手段が、

前記セルベクトルを参照して、前記テーブルにおいて各セルをクラスタリングしてセルクラスタ情報を作成するセルクラスタ作成手段と、

前記セル位置データおよび前記セルクラスタ情報を参照して、前記テーブル中のセルを結合してセグメントを生成するレイアウトセグメント生成手段とを備えたことを特徴とする請求項１に記載の文書分割装置。

【請求項５】　前記第１のセグメント生成手段が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項４に記載の文書分割装置。

【請求項６】　前記レイアウトセグメント生成手段が、前記テーブル中のセルの配置パターンを推定し、当該配置パターンに合致するセルを結合してセグメントを生成することを特徴とする請求項２に記載の文書分割装置。

【請求項７】　テーブルを1つのセグメントとして文書をセグメントに分割する一般セグメント生成手段を備え、

該一般セグメント生成手段により1つのセグメントとして生成されたテーブルを前記テーブル解析手段の処理対象とすることを特徴とする請求項１に記載の文書分割装置。

【請求項８】　処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、

該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、

前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、

前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とを備えたことを特徴とする文書分割方法。

【請求項９】　前記第１のセグメント生成工程が、

前記セル位置データおよび前記セルベクトルを参照して、前記テーブルにおいて各データが行または列のどちらで表現されているかを判別し、当該テーブルの分割方向を決める分割方向決定工程と、

前記テーブルタイプおよび前記分割方向を参照して、前記テーブルを分割してセグメントを生成する表セグメント生成工程とを備えたことを特徴とする請求項８に記載の文書分割方法。

【請求項１０】　前記第２のセグメント生成工程が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項９に記載の文書分割方法。

【請求項１１】　前記第２のセグメント生成工程が、

前記セルベクトルを参照して、前記テーブルにおいて各セルをクラスタリングしてセルクラスタ情報を作成するセルクラスタ作成工程と、

前記セル位置データおよび前記セルクラスタ情報を参照して、前記テーブル中のセルを結合してセグメントを生成するレイアウトセグメント生成工程とを備えたことを特徴とする請求項８に記載の文書分割方法。

【請求項１２】　前記第１のセグメント生成工程が、前記テーブルそのものをセグメントとして生成することを特徴とする請求項１１に記載の文書分割方法。

【請求項１３】　前記レイアウトセグメント生成工程が、前記テーブル中のセルの配置パターンを推定し、当該配置パターンに合致するセルを結合してセグメントを生成することを特徴とする請求項９に記載の文書分割方法。

【請求項１４】　テーブルを1つのセグメントとして文書をセグメントに分割する一般セグメント生成工程を備え、

該一般セグメント生成工程により1つのセグメントとして生成されたテーブルを前記テーブル解析工程の処理対象とすることを特徴とする請求項８に記載の文書分割方法。

【請求項１５】　処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、

該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、

前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、

前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とをコンピュータに実行させるための文書分割プログラムを記憶したことを特徴とする記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書を内容ごとに分割する文書分割装置とその方法、特に、テーブルを含む文書を分割する文書分割装置とその方法に関するものである。

【０００２】

【従来の技術】

従来、Web上の情報は「ページ」という単位で提供されており、ページの構成や大きさは情報提供者が自由に設定できる。もちろん、情報提供者は各自の情報伝達意図に基づいてページを作成しているのだが、それが必ずしも閲覧者の要求と一致しているとは限らない。

【０００３】

従って、情報提供者によって関連性が高いと判断された一連の話題が１ページにまとめられていても、閲覧者にとってはそれらの関連性が不要である可能性もあり、複数の話題のうちの１つだけが有用である場合には、他の話題の情報は必要な情報を探索する際の妨げにすらなる。特に、情報提示スペースの小さいモバイル機器では、必要な情報だけを表示するということが重要な機能となる。

【０００４】

そこで、表示対象である文書をあらかじめ内容ごとに分割しておき、閲覧者が必要としている部分だけを提示することが重要となる。Webページの大半は、Webページ記述言語であるHTML (Hyper Text Markup Language)を用いて書かれている。HTMLは文書構造を記述する言語であるが、論理構造の詳細を記述することは難しく、ブラウザにおけるレイアウトの指定が主な役割となっている。

【０００５】

しかし、ページのレイアウトには、情報提供者の情報に対する視点が反映されていると考えられる。そこで、情報提供者の意図を反映したセグメントを生成するために、HTMLのタグから読み取ったレイアウトに基づいてページを分割する手法が提案されている。

【０００６】

【発明が解決しようとする課題】

上記提案の手法では、<TABLE>タグと</TABLE>タグで囲まれたテーブルは、意味的なまとまりであると判断されて、１つのセグメントとして形成されている。しかしながら、テーブルは、比較的大きな領域を占めて複数の情報を含んでいる場合が多いため、さらに細かいセグメントに分割することが望ましい。

【０００７】

その際、テーブルは、単純な表を記述している場合と、テキストやイメージのレイアウトを指定している場合とがあるが、両者においてタグに含まれた意図はまったく異なるので、それぞれ違うアプローチでセグメントを生成すべきである。

【０００８】

単純な表を記述している場合は、含まれているデータごとにセグメントを生成することで、ユーザのより細かい要求に備えることができると考えられる。ところが、一口に表を記述していると言っても、１組のデータが行で表現されていたり列で表現されていたり、項目名を記述した行(または列)があったりなかったりと、様々な表の形式が存在する。従って、表をデータごとのセグメントに分割するためには、表の形式を判断する必要がある。

【０００９】

一方、テキストやイメージをレイアウトするためにテーブルタグを使っている場合は、各セルに記述された内容とセル同士の位置関係からセル間の関係を推定し、内容のまとまりを判断してセグメントを生成することが望まれる。

【００１０】

本発明は、上記の課題に鑑みてなされたものであり、処理対象となっているテーブルを解析して、表を記述したテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、文書中のテーブルを内容ごとに分割する文書分割装置を提供することを目的とする。

【００１１】

【課題を解決するための手段】

上述した目的を達成するために、本発明によれば、文書分割装置に、処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析手段と、該テーブル解析手段により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定手段と、前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成手段と、前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成手段とを備える。

【００１２】

また、本発明の他の態様によれば、文書分割方法に、処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とを備える。

【００１３】

更に、本発明の他の態様によれば、記憶媒体に、処理対象である文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析工程と、該テーブル解析工程により生成されたセル位置データおよびセルベクトルを参照してテーブルタイプを判定するテーブルタイプ判定工程と、前記テーブルタイプが表を記述したテーブルである場合に、前記テーブルからセグメントを生成する第１のセグメント生成工程と、前記テーブルタイプがレイアウトのためのテーブルである場合に、前記テーブルからセグメントを生成する第２のセグメント生成工程とをコンピュータに

実行させるための文書分割プログラムを記憶する。

【００１４】

【発明の実施の形態】

以下、図面を用いて本発明の１実施形態を詳細に説明する。

【００１５】

図１は、本実施形態の文書分割装置の機能構成を示すブロック図である。同図において、101は、処理対象であるHTML文書中のテーブル(<table>と</table>で囲まれた部分)を保持するHTMLテーブル保持部である。

【００１６】

102は、HTMLテーブル保持部101に保持されているテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成するテーブル解析部である。

【００１７】

セルベクトルは、セルの高さや幅、内容の表示位置、背景色、セル内のテキストの長さや文字種、セル内のイメージの大きさや形などから決定する。セルベクトルの次元は(セル内のイメージの個数×4＋17)次元であり、各成分は0以上1以下の実数である。

【００１８】

セル内でi番目に出現するイメージを$image_i$とするとき、セルベクトルvの第k成分v(k)は次のように定義される。

【００１９】

v(0)： タグの種類が<TH>(項目名を表現するセル)のとき1.0、<TD>(データを表現するセル)のとき0.0。

【００２０】

v(1)： rowspan（行幅）が4未満のときrowspan×0.25、4以上のとき1.0。

【００２１】

v(2)： colspan（列幅）が4未満のときcolspan×0.25、4以上のとき1.0。

【００２２】

v(3)： nowrap（改行なし）が指定されているとき1.0、指定されていないとき

0.0。

【００２３】

v(4)： align（横位置）の指定がないとき0.0、left（左詰め）のとき0.2、center（中央）のとき0.4、right（右詰め）のとき0.6、justify（均等）のとき0.8、それ以外のとき1.0。

【００２４】

v(5)： valign（縦位置）の指定がないとき0.0、top（上詰め）のとき0.2、middle（中央）のとき0.4、bottom（下詰め）のとき0.6、baseline（ベースライン）のとき0.8、それ以外のとき1.0。

【００２５】

v(6)： bgcolor（背景色）の指定がないとき0.0、16進コードで指定されていないとき0.0、16進コードで指定されているときbgcolor/0xFFFFFF。

【００２６】

v(7)： 9列目以前のとき(列番号)×0.1、10列目以降のとき1.0。

【００２７】

v(8)： 99行目以前のとき(行番号)×0.01、100行目以降のとき1.0。

【００２８】

v(9)： 改行(
)数が5つ未満のとき(改行数)×0.2、5つ以上のとき1.0。

【００２９】

v(10)： テキストの文字数が100文字未満のとき(文字数)×0.01、100文字以上のとき1.0。

【００３０】

v(11)： (テキスト中の数字の数)/(テキストの全文字数)。

【００３１】

v(12)： (テキスト中のアルファベットの数)/(テキストの全文字数)。

【００３２】

v(13)： (テキスト中の漢字の数)/(テキストの全文字数)。

【００３３】

v(14)： (テキスト中のカタカナの数)/(テキストの全文字数)。

【００３４】

v(15)：（テキスト中のひらがなの数)/(テキストの全文字数)。

【００３５】

v(16)：句点(“。”または“．”)があるとき1.0、ないとき0.0。

【００３６】

v(13＋i×4)：$image_i$の面積が150000未満のとき(面積)/150000、150000以上のとき1.0。

【００３７】

v(14＋i×4)：$image_i$の高さが300未満のとき(高さ)/300、300以上のとき1.0。

【００３８】

v(15＋i×4)：$image_i$の幅が500未満のとき(幅)/500、500以上のとき1.0。

【００３９】

v(16＋i×4)：このテーブルを含んでいるページのURLを示す文字列のうち、$image_i$のURLと共通の部分文字列の割合。例えば、
http://hogehoge.aaa.bbbbb.co.jp:8080/hoge1/hoge2/hoge.html
のページ(URLの長さは58)に“../image/hoge.gif”というイメージがあった場合、イメージをフルパスのURLに書き換えると、
http://hogehoge.aaa.bbbbb.co.jp:8080/hoge1/image/hoge.gif
となるので、共通の部分文字列は
http://hogehoge.aaa.bbbbb.co.jp:8080/hoge1/
となる。この長さは43なので、この成分の値は43÷58＝0.741となる。

【００４０】

103は、テーブル解析部102により生成されたセル位置データを保持するセル位置データ保持部である。104は、テーブル解析部102により生成されたセルベクトルを保持するセルベクトル保持部である。

【００４１】

105は、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照してテーブルタイプを判定し

、テーブルタイプによってカット方向決定部107、またはセルクラスタ作成部111に処理開始を指示するテーブルタイプ判定部である。テーブルタイプには、以下のtable I～table VIIの7種類がある。

【００４２】

table I： すべてのセルの高さと幅が1であり、1行n列目及びn行1列目のセルがすべて<TH>または同じ背景色。

【００４３】

table II： すべてのセルの高さと幅が1であり、1行n列目及びn行1列目(1行1列目を除く)のセルがすべて<TH>または同じ背景色。

【００４４】

table III： すべてのセルの高さと幅が1であり、1行n列目のセルがすべて<TH>または同じ背景色。

【００４５】

table IV： すべてのセルの高さと幅が1であり、1行n列目(1行1列目を除く)のセルがすべて<TH>または同じ背景色。

【００４６】

table V： すべてのセルの高さと幅が1であり、n行1列目のセルがすべて<TH>または同じ背景色。

【００４７】

table VI： すべてのセルの高さと幅が1であり、n行1列目(1行1列目を除く)のセルがすべて<TH>または同じ背景色。

【００４８】

table VII： table I～table VI以外のテーブル。

【００４９】

以上において、table I～table VIが表を記述するためのテーブルであり、table VIIがレイアウト目的のテーブルである。テーブルタイプがtable I～table VIの場合にはカット方向決定部107に処理開始を指示し、テーブルタイプがtable VIIの場合にはセルクラスタ作成部111に処理開始を指示する。

【００５０】

106は、テーブルタイプ判定部105により決定されたテーブルタイプを保持するテーブルタイプ保持部である。

【００５１】

107は、テーブルタイプ判定部105により処理開始を指示された場合に、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照して、表を記述したテーブルにおいて各データは行または列のどちらで表現されているかを判別し、テーブルの分割方向を決めるカット方向決定部である。

【００５２】

N行M列のテーブルTを行で分割したときのスコア$S_h(T)$と列で分割したときのスコア$S_v(T)$を以下のように定義する。以下で、$\cos(v_{i,j}, v_{k,l})$はi行j列目のセルのテーブルセルベクトル$v_{i,j}$とk行l列目のセルのテーブルセルベクトル$v_{k,l}$との余弦値を表す。

【００５３】

ただし、これはi行j列目のセルとk行l列目のセルのデータとが共に存在するときにのみ算出される値で、両方もしくはどちらか一方のセルのデータが存在しない場合には、値は0となる。

【００５４】

【外１】

$$\mathrm{exist}\ (i,j) = \begin{cases} 1 & (\text{i行j目列のセルにデータが存在する}) \\ 0 & (\text{i行j目列のセルにデータが存在しない}) \end{cases}$$

$$\mathrm{count}_h = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=j+1}^{M} \mathrm{exist}\ (i,j) \times \mathrm{exist}\ (i,k)$$

$$\mathrm{count}_v = \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{l=j+1}^{N} \mathrm{exist}\ (i,j) \times \mathrm{exist}\ (l,j)$$

$$S_h\ (T) = \frac{1}{\mathrm{count}_h} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=j+1}^{M} \cos\ (v_{ij}, v_{ik})$$

$$S_v\ (T) = \frac{1}{\mathrm{count}_v} \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{l=j+1}^{N} \cos\ (v_{ij}, v_{lj})$$

【００５５】

テーブルセルベクトルの次元は、i行j列目のセルとk行l列目のセルに含まれるイメージの数により決定されるので、両ベクトルの次元が同じになるように、低次元のテーブルセルベクトルに値0の成分を追加して余弦値を計算する。

【００５６】

$S_h(T)$は同じ行にある２つのセルのテーブルセルベクトルの平均余弦値であり、$S_v(T)$は同じ列にある２つのセルのテーブルセルベクトルの平均余弦値である。２つのテーブルセルベクトルの余弦値はセルの類似度と見なせるので、$S_h(T)$はテーブルを行ごとに分割した時の同セグメント内におけるセル間の平均類似度、$S_v(T)$はテーブルを列ごとに分割した時の同セグメント内におけるセル間の平均類似度といえる。

【００５７】

各セグメントに各種のデータを盛り込むには、同セグメント内セル間類似度が低い方が良いので、$S_h(T) \leqq S_v(T)$のときはテーブルTを行ごとに分割し、$S_h(T) > S_v(T)$のときテーブルTを列ごとに分割するべきだと判断する。

【００５８】

108は、カット方向決定部107により決定されたカット方向を保持するカット方

向保持部である。

【００５９】

109は、テーブルタイプ保持部106に保持されたテーブルタイプ、およびカット方向保持部108に保持されたカット方向を参照して、表を記述したテーブルからセグメントを生成する表セグメント生成部である。カット方向が行方向の場合、table Vのテーブルはそのまま行をセグメントとし、table V以外のテーブルは1行目を組み合わせてセグメントを作る。カット方向が列方向の場合、table IIIのテーブルはそのまま列をセグメントとし、table III以外のテーブルは1列目を組み合わせてセグメントを作る。

【００６０】

110は、表セグメント生成部109により生成された表セグメントを保持する表セグメント保持部である。

【００６１】

111は、テーブルタイプ判定部105により処理開始を指示された場合に、セルベクトル保持部104に保持されたセルベクトルを参照して、レイアウト目的のテーブルにおいて各セルをクラスタリングするセルクラスタ作成部である。ここでは最大距離アルゴリズムを用いてセルの分類を決定する。最大距離アルゴリズムのクラスタリング手順を以下に示す。

【００６２】

Step.1： N個のサンプルパターン集合$X = \{x_1, x_2, \cdots, x_N\}$ から、任意にひとつ(ここではx_1として説明する)を選び、クラスタ中心$z_1 \in Z$とする。

【００６３】

Step.2： Zに含まれないすべての$x_i \in X$について、すでに選ばれたクラスタ中心$z_j \in Z$のうち、一番近いものまでの距離dx_iを計算する。$\mathrm{Max}\{dx_i\}$を与えるx_iをx_Cとする。

【００６４】

Step.3： すべての$z_k \in Z$について、z_k以外のクラスタ中心のうち、一番遠いものまでの距離dz_kを計算する。

【００６５】

Step.4： $dx_C \geqq \max\{dz_k\} \times t(t=0.5\sim1)$が成立するとき、$x_C$を新たなクラスタ中心とし、Step.2に戻って次のクラスタ中心を選ぶ。$dx_C < \max\{dz_k\} \times t(t=0.5\sim1)$ならばStep.5へ。

【００６６】

Step.5： すべての$x_i \in X$を、最も近い$z_j \in Z$のクラスタに分類する。

【００６７】

最大距離アルゴリズムによるクラスタリング結果の例を図４に示す。

【００６８】

112は、セルクラスタ作成部111により作成されたセルのクラスタ情報を保持するセルクラスタ情報保持部である。

【００６９】

113は、セル位置データ保持部103に保持されたセル位置データ、およびセルクラスタ情報保持部112に保持されたセルクラスタ情報を参照して、レイアウト目的のテーブルからセグメントを生成するレイアウトセグメント生成部である。

【００７０】

テーブルの形式を利用して情報を配置するメリットとしては、ある配置パターンの縦横方向の繰り返しが容易に表現できる点が挙げられる。そこで、セルクラスタ情報をもとに配置パターンを推定して、パターンに適合するセルを合わせてセグメントとする。ある配置パターンが繰り返し現れるときには、そのパターンに適合するセル同士が意味的にまとまっていると判断できるからである。処理の詳細を以下に示す。

【００７１】

まず、基本セル種を決定し、基本セル種に属するセルを基本セルとする。基本セル種は、同種のセルが複数あるセルの種類のうち、最もセル数の少ないセル種とする。該当するセル種が複数ある場合には、より左、上にあるセルの種類を選ぶ。

【００７２】

次に、ある基本セルに隣接するセルと分類が同じセルが他の基本セルにも同じ

ように隣接するかを確認する。隣接していれば、それぞれを結合し、新たな基本セルとする。これを結合できなくなるまで繰り返す。

【００７３】

以上の処理を終えると、基本セルおよび残りのセルをそれぞれセグメントとする。

【００７４】

114は、レイアウトセグメント生成部113により生成されたレイアウトセグメントを保持するレイアウトセグメント保持部である。表セグメント保持部110に保持された表セグメント、およびレイアウトセグメント保持部114に保持されたレイアウトセグメントが結果として得られるセグメントである。

【００７５】

図２は、本発明の実施形態に係る文書分割装置のハードウェア構成を示す図である。

【００７６】

同図において、201は後述する制御手順を実現するプログラムを保持するROMである。202はRAMで、セル位置データ保持部103、セルベクトル保持部104、テーブルタイプ保持部106、カット方向保持部108、セルクラスタ情報保持部112と上記プログラムの動作に必要な記憶領域とを提供する。

【００７７】

203はROM201に保持されているプログラムに従って処理を行なう中央処理装置である。204はディスク装置であり、HTMLテーブル保持部101、表セグメント保持部110、レイアウトセグメント保持部114を実現する。205はバスであり、上記の各構成を接続し、各構成間におけるデータの授受を可能とする。

【００７８】

次に、本実施形態の処理動作を説明する。図３は本実施形態の文書分割装置の動作手順を示すフローチャートである。

【００７９】

ステップS301では、HTMLテーブル保持部101に保持されているテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセル

ベクトルを生成する。そしてステップS302に移る。

【００８０】

ステップS302では、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照して、テーブルタイプを決定する。そしてステップS303に移る。

【００８１】

ステップS303では、テーブルタイプ保持部106に保持されたテーブルタイプを参照して、処理対象のテーブルが表を記述したテーブルか否かを判定して、表を記述したテーブルの場合はステップS304に移る。表を記述したテーブルでない場合はステップS306に移る。

【００８２】

ステップS304では、セル位置データ保持部103に保持されたセル位置データ、およびセルベクトル保持部104に保持されたセルベクトルを参照して、表を記述したテーブルにおいて各データは行または列のどちらで表現されているかを判別し、テーブルの分割方向を決める。そしてステップS305に移る。

【００８３】

ステップS305では、テーブルタイプ保持部106に保持されたテーブルタイプ、およびカット方向保持部108に保持されたカット方向を参照して、表を記述したテーブルからセグメントを生成する。そして動作を終了する。

【００８４】

ステップS306では、セルベクトル保持部104に保持されたセルベクトルを参照して、レイアウト目的のテーブルにおいて各セルをクラスタリングする。そしてステップS307に移る。

【００８５】

ステップS307では、セル位置データ保持部103に保持されたセル位置データ、およびセルクラスタ情報保持部112に保持されたセルクラスタ情報を参照して、レイアウト目的のテーブルからセグメントを生成する。そして動作を終了する。

【００８６】

以上に述べたように、処理対象となっているテーブルを解析して、表を記述し

たテーブルであるか、レイアウト目的のテーブルであるかを判別し、それぞれに応じた処理によってセグメントを生成することで、HTML文書中のテーブルを内容ごとに分割する文書分割装置を実現することができる。

【００８７】

〔他の実施形態〕

上記実施形態では、セルのクラスタリングに最大距離アルゴリズムを利用するよう説明しているが、これに限定されるものではなく、他のアルゴリズムを用いてクラスタリングを行なってもよい。

【００８８】

上記実施形態で示したセルベクトルの各成分の定義は一例であり、他の定義によってセルの特徴をベクトル表現してもよい。

【００８９】

上記実施形態で示したカット方向を決定するスコアの定義は一例であり、他の定義によってカット方向を決定してもよい。

【００９０】

上記実施形態では、テーブルタイプを決定するための項目名の行(または列)の判定に、セルの高さと幅、タグの種類(TH or TD)、背景色を用いているが、これに限定されるものではなく、他の属性を用いて判定してもよい。

【００９１】

上記実施形態では、HTMLのテーブルを分割するだけの装置として説明しているが、これに限定されるものではない。例えば、HTML文書全体を分割する装置であってもよい。図５は、この場合の基本的な機能構成を示すブロック図である。

【００９２】

図５において、501は、処理対象であるHTML文書を保持するHTML文書保持部である。502は、HTML文書保持部501に保持されているHTML文書をセグメントに分割する一般セグメント生成部である。503は、一般セグメント生成部502により生成されたテーブル以外のセグメントを保持する一般セグメント保持部である。504は、一般セグメント生成部502により生成されたテーブルのセグメントを保持するHTMLテーブル保持部である。

【００９３】

以下、505～517は、図1の102～114と同様である。

【００９４】

図5では、一般セグメント保持部503に保持された一般セグメント、表セグメント保持部513に保持された表セグメント、およびレイアウトセグメント保持部517に保持されたレイアウトセグメントが結果として得られるセグメントである。

【００９５】

上記実施形態では、表を記述しているテーブルとレイアウト目的のテーブルの両方をセグメントに分割しているが、これに限定されるものではない。例えば、表を記述しているテーブルのみを分割してもよい。図６はこの場合の基本的な機能構成を示すブロック図である。

【００９６】

図６において、601～610は、図1の101～110と同様である。

【００９７】

611は、テーブルタイプ判定部605により処理開始を指示された場合に、HTMLテーブル保持部601に保持されたHTMLテーブルをテーブルセグメントとするテーブルセグメント生成部である。

【００９８】

612は、テーブルセグメント生成部611により生成されたテーブルセグメントを保持するテーブルセグメント保持部である。

【００９９】

図6では、表セグメント保持部610に保持された表セグメント、およびテーブルセグメント保持部612に保持されたテーブルセグメントが結果として得られるセグメントである。

【０１００】

また、上記実施形態では、表を記述しているテーブルとレイアウト目的のテーブルの両方をセグメントに分割しているが、レイアウト目的のテーブルのみを分割してもよい。図７はこの場合の基本的な機能構成を示すブロック図である。

【０１０１】

図７において、701～705及び708～711は、図1の101～115及び111～114と同様である。

【０１０２】

706は、テーブルタイプ判定部705により処理開始を指示された場合に、HTMLテーブル保持部701に保持されたHTMLテーブルをテーブルセグメントとするテーブルセグメント生成部である。707は、テーブルセグメント生成部706により生成されたテーブルセグメントを保持するテーブルセグメント保持部である。

【０１０３】

図7では、テーブルセグメント保持部707に保持されたテーブルセグメント、およびレイアウトセグメント保持部711に保持されたレイアウトセグメントが、結果として得られるセグメントである。

【０１０４】

上記実施形態では、HTML文書を分割する装置として説明しているが、これに限定されるものではなく、検索装置と組み合わせて、生成されたセグメント単位で検索を行なうことができるセグメント検索装置として実現してもよい。

【０１０５】

上記実施形態においては、セル位置データ保持部103、セルベクトル保持部104、テーブルタイプ保持部106、カット方向保持部108、セルクラスタ情報保持部112をRAMで、HTMLテーブル保持部101、表セグメント保持部110、レイアウトセグメント保持部114をディスク装置で実現する場合について説明したが、これに限定されるものではなく、任意の記憶媒体を用いて実現してもよい。

【０１０６】

上記実施形態では、HTMLのテーブルを分割する場合について説明したが、テーブルの内容が区別できれば、他の形式であってもよい。

【０１０７】

上記実施形態においては、各部を同一の計算機上で構成する場合について説明したが、これに限定されるものではなく、ネットワーク上に分散した計算機や処理装置などに分かれて各部を構成してもよい。

【０１０８】

上記実施形態においては、プログラムをROMに保持する場合について説明したが、これに限定されるものではなく、任意の記憶媒体を用いて実現してもよい。また、同様の動作をする回路で実現してもよい。

【０１０９】

なお、本発明は、複数の機器から構成されるシステムに適用しても、１つの機器からなる装置に適用してもよい。前述した実施形態の機能を実現するソフトウエアのプログラムコードを記録した記録媒体を、システム或いは装置に供給し、そのシステム或いは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記録媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。

【０１１０】

この場合、記録媒体から読み出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記録した記録媒体は本発明を構成することになる。

【０１１１】

プログラムコードを供給するための記録媒体としては、例えば、フロッピーディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【０１１２】

また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【０１１３】

更に、記録媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボ

ードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【０１１４】

【発明の効果】

以上説明したように、本発明によれば、文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと、各セルの特徴を表現したセルベクトルとを生成し、このセル位置データおよびセルベクトルを参照して、処理対象のテーブルが表を記述したテーブルか否かを判定し、判定結果に応じた手法でセグメントを生成することで、文書中のテーブルを内容ごとに分割する文書分割を実現できるという効果が得られる。

【図面の簡単な説明】

【図１】

本発明に係る一実施形態の文書分割装置の基本構成を示すブロック図である。

【図２】

実施形態に係る文書分割装置のハードウエア構成を示すブロック図である。

【図３】

実施形態に係る文書分割装置の動作手順を示すフローチャートである。

【図４】

最大距離アルゴリズムを説明する図である。

【図５】

他の実施形態の基本構成を示すブロック図である。

【図６】

他の実施形態の基本構成を示すブロック図である。

【図７】

他の実施形態の基本構成を示すブロック図である。

【符号の説明】

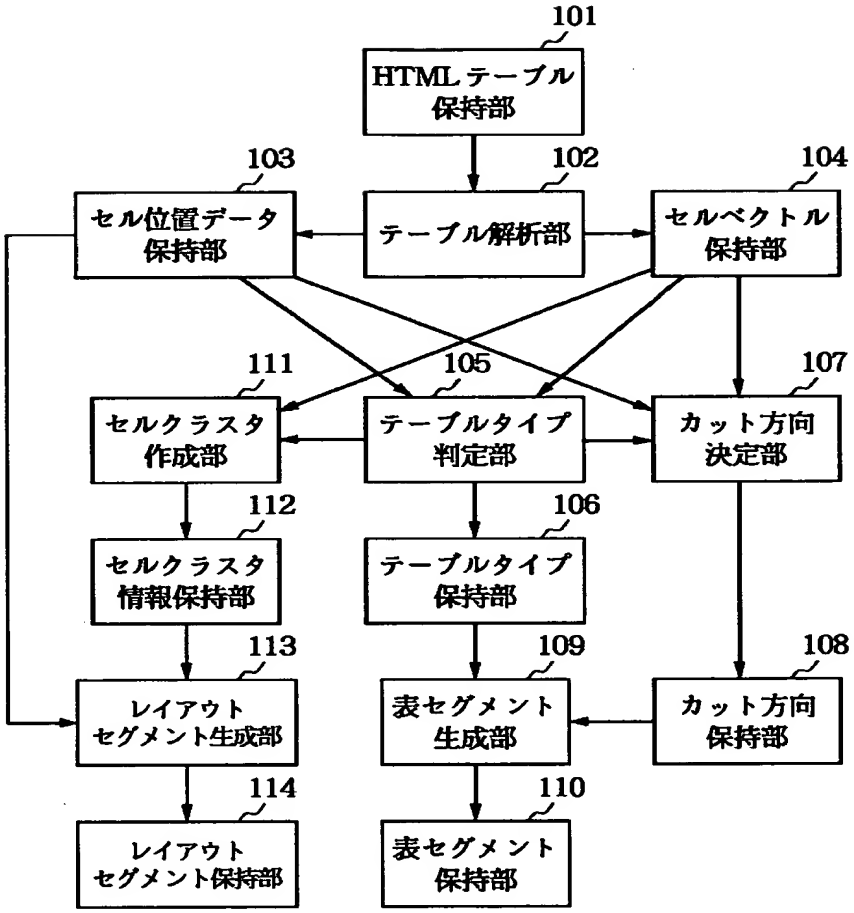
１０１　ＨＴＭＬテーブル保持部

１０２　テーブル解析部

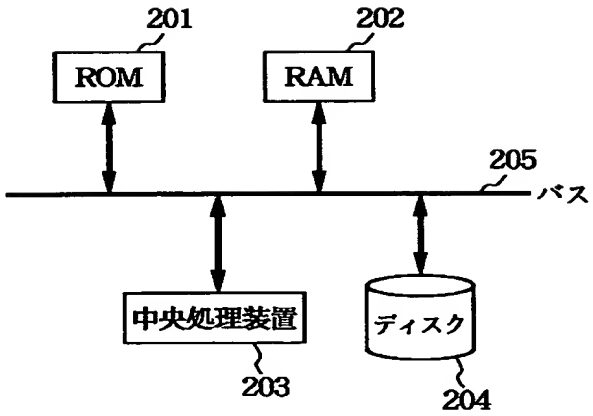
１０３　セル位置データ保持部
１０４　セルベクトル保持部
１０５　テーブルタイプ判定部
１０６　テーブルタイプ保持部
１０７　カット方向決定部
１０８　カット方向保持部
１０９　表セグメント生成部
１１０　表セグメント保持部
１１１　セルクラスタ作成部
１１２　セルクラスタ情報保持部
１１３　レイアウトセグメント生成部
１１４　レイアウトセグメント保持部
２０１　ＲＯＭ
２０２　ＲＡＭ
２０３　中央処理装置
２０４　ディスク
２０５　バス

【書類名】　　図面

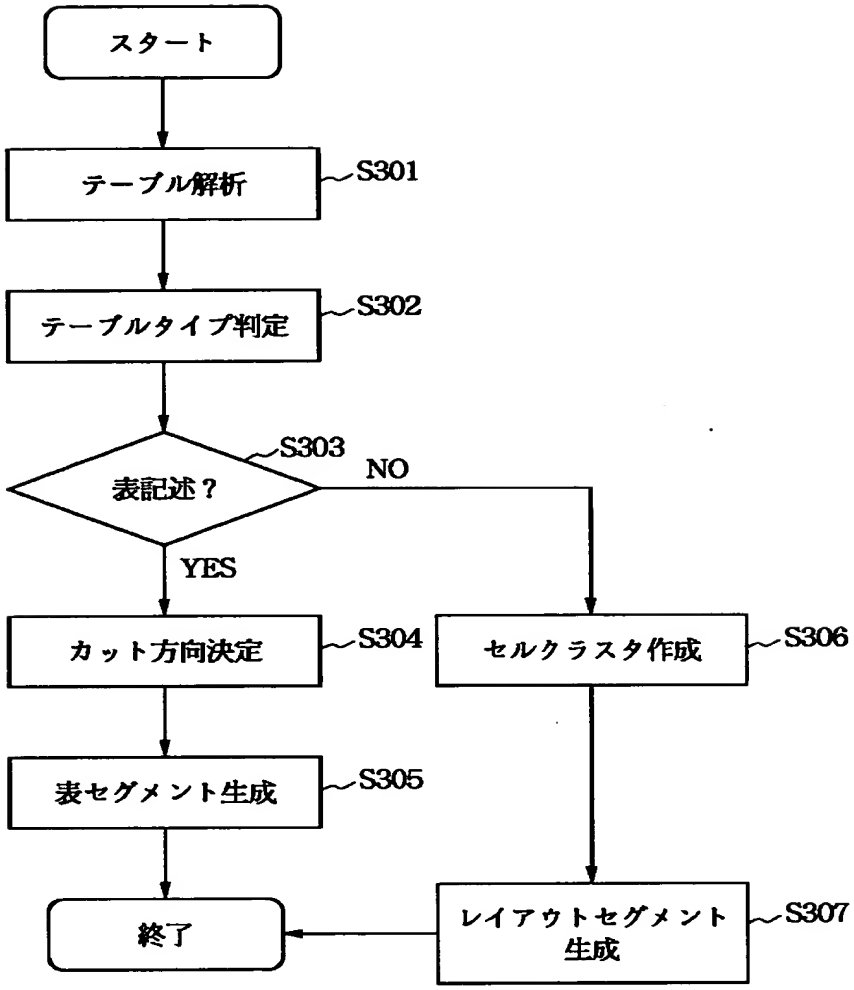
【図１】


101
HTMLテーブル保持部
103
セル位置データ保持部
102
テーブル解析部
104
セルベクトル保持部
111
セルクラスタ作成部
105
テーブルタイプ判定部
107
カット方向決定部
112
セルクラスタ情報保持部
106
テーブルタイプ保持部
113
レイアウトセグメント生成部
109
表セグメント生成部
108
カット方向保持部
114
レイアウトセグメント保持部
110
表セグメント保持部

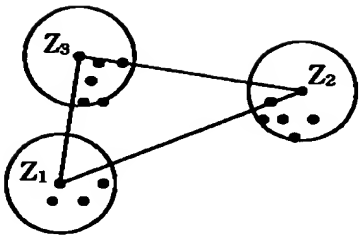
【図２】



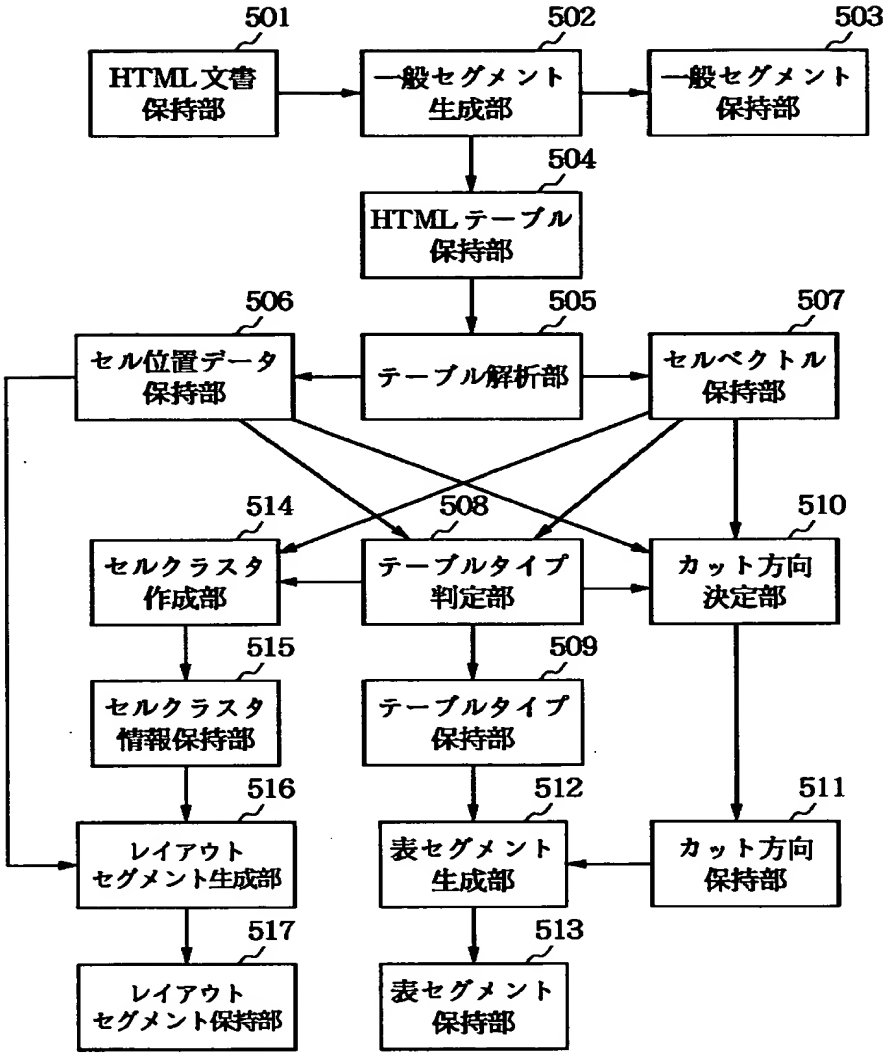
【図３】


スタート
テーブル解析
S301
テーブルタイプ判定
S302
表記述？
S303
NO
YES
カット方向決定
S304
セルクラスタ作成
S306
表セグメント生成
S305
終了
レイアウトセグメント生成
S307

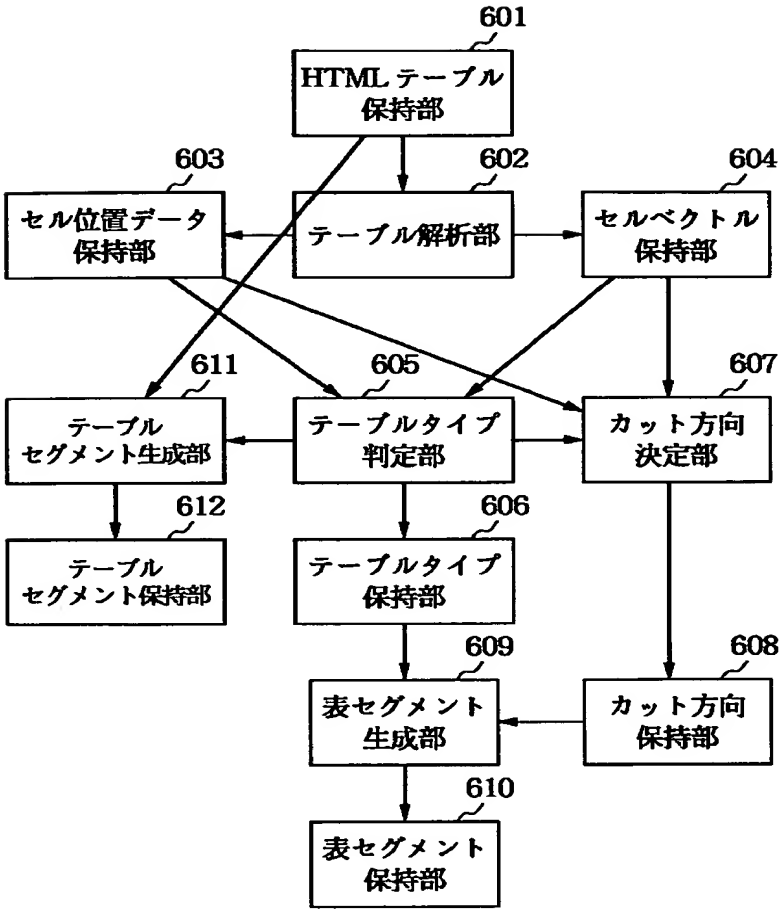
【図４】



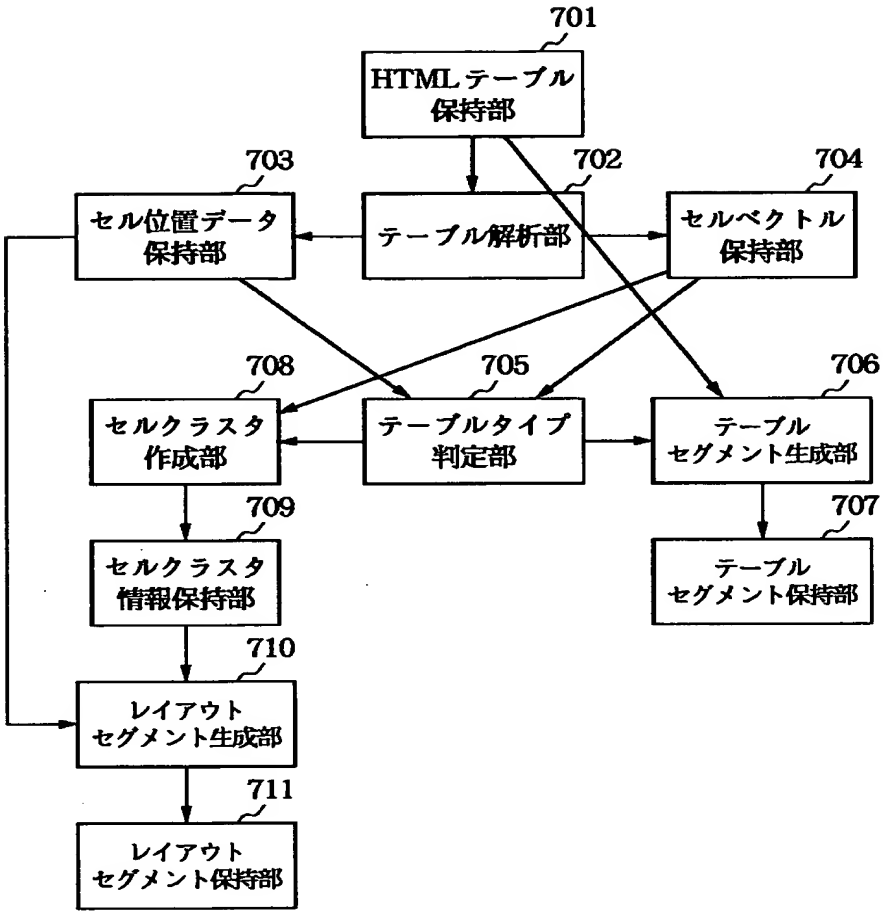
【図５】


501
HTML文書保持部
502
一般セグメント生成部
503
一般セグメント保持部
504
HTMLテーブル保持部
506
セル位置データ保持部
505
テーブル解析部
507
セルベクトル保持部
514
セルクラスタ作成部
508
テーブルタイプ判定部
510
カット方向決定部
515
セルクラスタ情報保持部
509
テーブルタイプ保持部
516
レイアウトセグメント生成部
512
表セグメント生成部
511
カット方向保持部
517
レイアウトセグメント保持部
513
表セグメント保持部

【図６】


601
HTMLテーブル保持部
603
セル位置データ保持部
602
テーブル解析部
604
セルベクトル保持部
611
テーブルセグメント生成部
605
テーブルタイプ判定部
607
カット方向決定部
612
テーブルセグメント保持部
606
テーブルタイプ保持部
609
表セグメント生成部
608
カット方向保持部
610
表セグメント保持部

【図７】


701
HTMLテーブル保持部
703
セル位置データ保持部
702
テーブル解析部
704
セルベクトル保持部
708
セルクラスタ作成部
705
テーブルタイプ判定部
706
テーブルセグメント生成部
709
セルクラスタ情報保持部
707
テーブルセグメント保持部
710
レイアウトセグメント生成部
711
レイアウトセグメント保持部

【書類名】　　要約書

【要約】

【課題】　HTML文書中のテーブルを内容ごとに分割する。

【解決手段】　HTML文書中のテーブルを解析して、各セルの位置関係を示すセル位置データと各セルの特徴を表現したセルベクトルとを生成し(S301)、このセル位置データおよびセルベクトルを参照してテーブルタイプを判定し(S302)、表を記述したテーブルの場合は、セル位置データおよびセルベクトルを参照して、各データは行または列のどちらで表現されているかを判別し、テーブルの分割方向を決め(S304)、テーブルタイプおよび分割方向を参照してセグメントを生成し(S305)、表を記述したテーブルでないレイアウト目的のテーブルの場合は、セルベクトルを参照して各セルをクラスタリングし(S306)、セル位置データおよびセルクラスタ情報を参照してセグメントを生成する(S307)。

【選択図】　　図３

出　願　人　履　歴　情　報

識別番号　　〔０００００１００７〕

１．変更年月日　　１９９０年　８月３０日
［変更理由］　新規登録
住　所　　東京都大田区下丸子３丁目３０番２号
氏　名　　キヤノン株式会社